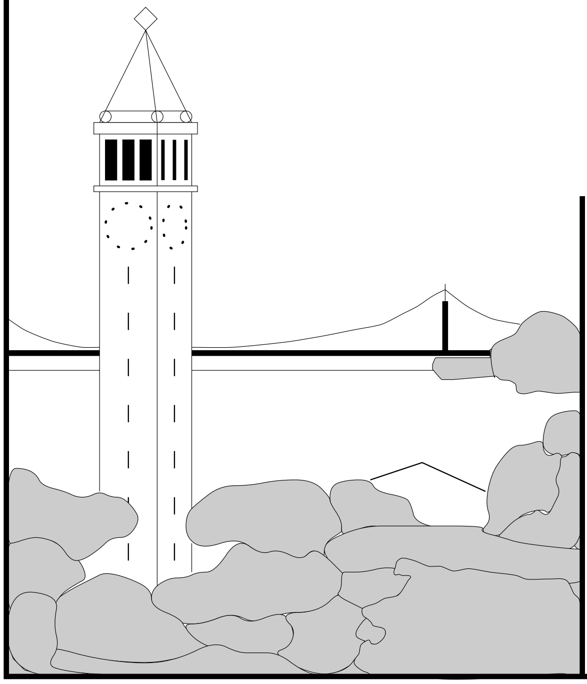
Multi-Paragraph Segmentation of Expository Texts

Marti A. Hearst



Report No. UCB/CSD 94/790

January 1994

Computer Science Division (EECS)
University of California
Berkeley, California 94720

UC Berkeley Computer Science Technical Report Number UCB/CSD-94-790

Multi-Paragraph Segmentation of Expository Texts

Marti A. Hearst
Computer Science Division, 571 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
and
Xerox Palo Alto Research Center
marti@cs.berkeley.edu[*]

Abstract

We present a method for partitioning expository texts into coherent multi-paragraph units which reflect the subtopic structure of the texts. Using Chafe's Flow Model of discourse, we observe that subtopics are often expressed by the interaction of multiple simultaneous themes. We describe two fully-implemented algorithms that use only term repetition information to determine the extents of the subtopics. We show that the segments correspond well to human judgements of the major subtopic boundaries of thirteen lengthy texts, and suggest the use of such segments in information retrieval applications.

1 Introduction: Multi-paragraph Segmentation

The structure of expository texts can be characterized as a sequence of subtopical discussions that occur in the context of a few main topic discussions. For example, a text called *Stargazers*, whose main topic is the existence of life on earth and other planets, could be judged to consist of the following subdiscussions (numbers indicate paragraph numbers):

1-3	*Intro – the search for life in space*
4-5	*The moon's chemical composition*
6-8	*How ancient nearness of the moon shaped it*
9-12	*How the moon helped life evolve on earth*
13	*The improbability of the earth-moon system*
14-16	*Binary/trinary star systems make life unlikely*
17-18	*The low probability of non-binary/trinary systems*
19-20	*Properties of our sun that facilitate life*
21	*Summary*

[*]This research was sponsored in part by the Advanced Research Projects Agency under Grant No. MDA972-92-J-1029 with the Corporation for National Research Initiatives (CNRI).

Subtopic structure is sometimes marked in technical texts by headings and subheadings which divide the text into coherent segments; (Brown & Yule 1983) (p 140) state that this kind of division is one of the most basic in discourse. However, many expository texts consist of long sequences of paragraphs with very little structural demarcation. In this paper we present fully-implemented algorithms that use lexical cohesion relations to partition expository texts into multi-paragraph segments that reflect their subtopic structure. In essence, our goal is to let the meaning of the text determine its structure. We introduce the use of multiple simultaneous themes as an indicator of subtopical extent and we compare the results of two segmentation algorithms against an upper bound of reader judgements and a lower bound baseline measurement.

Most discourse segmentation is done at a finer granularity than that suggested here. However, for lengthy written expository texts, multi-paragraph segmentation has many potential uses, including the improvement of computational tasks that make use of distributional information. For example, disambiguation algorithms that train on arbitrary-size text windows (e.g., (Yarowsky 1992), (Gale *et al.* 1992b)) and algorithms that use lexical co-occurrence to determine semantic relatedness (e.g., (Schütze 1993)) might benefit from using windows with motivated boundaries instead.

Information retrieval algorithms can use subtopic structuring to return meaningful portions of a text if paragraphs are too short and sections are too long (or are not present). Motivated segments can also be used as a more meaningful unit for indexing long texts. (Salton *et al.* 1993), working with encyclopedia text, find that comparing a query against sections and then paragraphs is more successful than comparing against full documents alone. They also discovered (personal communication) that comparing first against sections and then against paragraphs worked better than comparing against paragraphs alone in the cases in which adjacent paragraphs should have been grouped together because their content was similar. We have preliminary results, described in (Hearst & Plaunt 1993), that indicate that multi-paragraph segmentation helps improve results like these, especially in texts less well-behaved than encyclopedia articles.

Finally, (Mooney *et al.* 1990) have found multi-paragraph units useful for text generation, implying that this unit of segmentation should be useful for recognition tasks as well.

In what follows, Section 2 describes the discourse model that motivates our approach, Section 3 describes two algorithms for subtopic structuring that make use only of lexical cohesion relations, Section 4 presents the evaluation of these algorithms, and Section 5 discusses future work and summarizes the paper.

2 The Discourse Model

Many discourse models assume a hierarchical segmentation model (e.g., attentional/intentional structure (Grosz & Sidner 1986), Rhetorical Structure Theory (Mann & Thompson 1987)). Although many aspects of discourse analysis require such a model, we choose to cast expository text into a linear sequence of segments, both for computational simplicity and because such a structure is sufficient for the coarse-grained tasks we are pursuing.[1] In doing this, we are influenced by (Skorochod'ko 1972), who suggests determining the semantic structure of a text (for the purposes of automatic abstracting) by analyzing it in terms of the topology formed by semantic interrelations found among its sentences, as shown in Figure 1.

Since our goal is to identify sequences of subtopical discussions, we are concerned principally with the last topology; it represents sequences of densely interrelated discussions linked together.

Figure 1 illustrates the important point that a subtopical discussion consists of several

[1] Additionally, (Passonneau & Litman 1993) concede the difficulty of eliciting hierarchical intentional structure with any degree of consistency from their human judges.

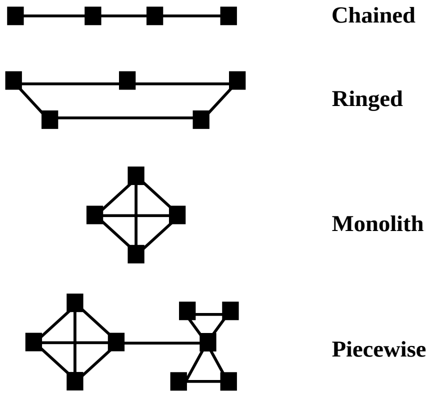


Figure 1: Skorodch'ko's text structure types. Nodes correspond to sentences and edges between nodes indicate strong term overlap between the sentences.

different themes discussed simultaneously. This theoretical stance bears a close resemblance to Chafe's notion of The Flow Model of discourse (Chafe 1979), in description of which he writes (pp 179-180):

> Our data ... suggest that as a speaker moves from focus to focus (or from thought to thought) there are certain points at which there may be a more or less radical change in space, time, character configuration, event structure, or, even, world. ... At points where all of these change in a maximal way, an episode boundary is strongly present. But often one or another will change considerably while others will change less radically, and all kinds of varied interactions between these several factors are possible.[2]

Although Chafe's work concerns narrative text, we feel the same kind of observation applies to expository text. Our algorithms are designed to recognize episode boundaries by determining where the thematic components listed by Chafe change in a maximal way.

Many researchers have studied the patterns of occurrence of characters, setting, time, and the other thematic factors that Chafe mentions, usually in the context of narrative. In contrast, we attempt to determine where a relatively large set of active themes changes simultaneously, regardless of the *type* of thematic factor. This is especially important in expository text in which the subject matter tends to structure the discourse more so than characters, setting, etc.[3] For example, in the *Stargazers* text, a discussion of continental movement, shoreline acreage, and habitability gives way to a discussion of binary and unary star systems. This is not so much a change in setting or character as a change in subject matter.

Therefore, to recognize where the subtopic changes occur, we make use of lexical cohesion relations (Halliday & Hasan 1976) in a manner similar to that suggested by Skorodch'ko.

[2] Interestingly, Chafe arrived at the Flow Model after working extensively with, and then becoming dissatisfied with, a Longacre-style hierarchical model of paragraph structure (Longacre 1979).

[3] cf. (Sibun 1992) for a discussion of how the form of people's descriptions often mirror the form of what they are describing.

This differs from the work of (Morris & Hirst 1991) in several ways, the most important of which is that our algorithms emphasize the interaction of multiple simultaneous themes, rather than following single threads of discussion alone. Main topics are themes that continue on throughout the ebb and flow of the interacting subtopics.

(Morris & Hirst 1991) use lexical cohesion relations of semantically related terms, via a thesaurus, to find Grosz and Sidner-style intentional structure in short texts at the phrase level. Their algorithm finds chains of related terms; the extent of the chains correspond to the extent of a segment. They also incorporate the notion of "chain returns" – repetition of terms after a long hiatus – to close off an intention that spans over a digression. Bearing in mind that the goal of their algorithm is to find finer-grained distinctions than those desired here, the model is not set up to take advantage of the fact that multiple simultaneous chains might occur over the same intention. For example, in Text 4-3 of (Morris 1988), a discussion of the role of women in the USSR as embodied in the life of Raisa Gorbachev, two different chains span most of the text: One consists of terms relating to the Soviet Union and the United States, and the other refers to women, men, husbands, and wives.

Furthermore, chains tend to overlap one another extensively in long texts. Figure 2 shows the distribution, by sentence number, of selected terms from the *Stargazers* text. The first two terms have fairly uniform distribution and so should not be expected to provide much information about the divisions of the discussion. The next two terms occur mainly at the beginning of the text, while terms *binary* through *planet* have considerable overlap from sentences 58 to 78. There is a somewhat well-demarked cluster of terms between sentences 35 and 50, corresponding to the grouping together of paragraphs 10, 11, and 12 by human judges who have read the text.

From the diagram it is evident that simply looking for chains of repeated terms is not sufficient for determining subtopic breaks. Even combining terms that are closely related semantically into single chains is insufficient, since often several different themes are active in the same segment. For example, sentences 37 - 51 contain dense interaction among the terms *move, continent, shoreline, time, species,* and *life*, and all but the latter occur only in this region. Few thesauri would group all of these terms together. However, it is the case that the interlinked terms of sentences 57 - 71 (*space, star, binary, trinary, astronomer,orbit*) are closely related semantically, assuming the appropriate senses of the terms have been determined.

3 Algorithms for Discovering Subtopic Structure

Many researchers (e.g., (Halliday & Hasan 1976), (Tannen 1989), (Walker 1991)) have noted that term repetition is a strong cohesion indicator. We have found in this work that term repetition alone, when used in terms of multiple simultaneous threads of information, are a very useful indicator of subtopic structure. This section describes two algorithms for discovering subtopic structure using term repetition as a lexical cohesion indicator as an approximation to Chafe's Flow Model.

The first method compares, for a given window size, each pair of adjacent blocks of text according to how similar they are lexically. This method assumes that the more similar two blocks of text are, the more likely it is that the current subtopic continues, and, conversely, if two adjacent blocks of text are dissimilar, this implies a change in subtopic flow. The second method, an extension of Morris's approach (Morris & Hirst 1991), keeps track of active chains of repeated terms, where membership in a chain is determined by location in the text. The method determines subtopic flow by recording where in the discourse the bulk of one set of chains ends and a new set of chains begins. Figure 3 illustrates the two approaches.

Through experimentation we discovered best results for the block similarity algorithm are obtained with a block size of 6 sentences. In our previous experiments we weighted the terms by their frequency in a block times the inverse frequency of the number of blocks

```
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
14       form    1     111 1   1                               1 1     1   1       1     1     1    1
 8  scientist                   11            1   1               1          1      1  1
 5      space 11    1     1                                                            1
25       star    1             1                                       11 22  111112  1 1  1   11 1111     1
 5     binary                                                           11  1          1                    1
 4    trinary                                                            1   1         1                    1
 8 astronomer 1                1                                        1 1           1   1    1  1
 7      orbit   1                    1                                     12    1 1
 6       pull                          2     1 1                               1  1
16     planet   1    1       11               1           1                21  11111                  1     1
 7     galaxy    1                                                   1                1  11     1           1
 4      lunar            1  1     1       1
19       life 1  1  1                              1    11 1  11  1      1                 1 1    1 111  1 1
27       moon        13  1111   1 1 22 21  21    21           11 1
 3       move                                       1   1   1
 7  continent                                       2 1 1 2 1
 3  shoreline                                             12
 6       time                      1               1  1  1     1                                          1
 3      water                                  11            1
 6        say                                 1 1        1        11                1
 3    species                                         1  1  1
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
```

Figure 2: Distribution of selected terms from the *Stargazer* text, with a single digit frequency per sentence number (blanks indicate a frequency of zero).

they appear in. In these more recent experiments we found that weighting the terms by simple frequency works better. To compute similarity between blocks, term counts are placed in vectors which are compared using the cosine distance measure, yielding a score between 0 and 1. The best variation on the chaining algorithm allows gaps of up to six sentences before the chain is considered to be broken. For both algorithms, sentences are actually blocks of 20 words (ignoring punctuation) rather than real sentence boundaries, in order to avoid normalization problems. Additionally, morphological analysis (inflectional) is used to normalize different word forms and a stop list of 898 words is used to eliminate common words from the calculations.

The results of both algorithms are plotted as similarity against sentence gap number. Boundaries are scored according to the relative depths of the valleys in the resulting plots, thus breaks in similarity adjacent to high strong peaks (indicating dense cohesion relations) are considered stronger boundaries than those near lesser peaks. The actual values of the similarity measures are not taken into account; the relative differences are what are of consequence. The valley depth must exceed a threshold; by experimentation we devised one that scales with the size of the text and is a function of the average and standard deviations of the valley depths for each text. The block algorithm works best with one iteration of average smoothing with a window size of 3; chaining works best with no smoothing.

4 Evaluation

One way to evaluate these segmentation algorithms is to compare against judgements made by human readers, which we do here. Another is to see how well the results improve a computational task; we report preliminary work in this vein in (Hearst 1994). A third possible evaluation – comparing the algorithms against texts pre-marked by authors – is not possible since the kinds of texts we are interested in working on by definition are not pre-marked in this way.

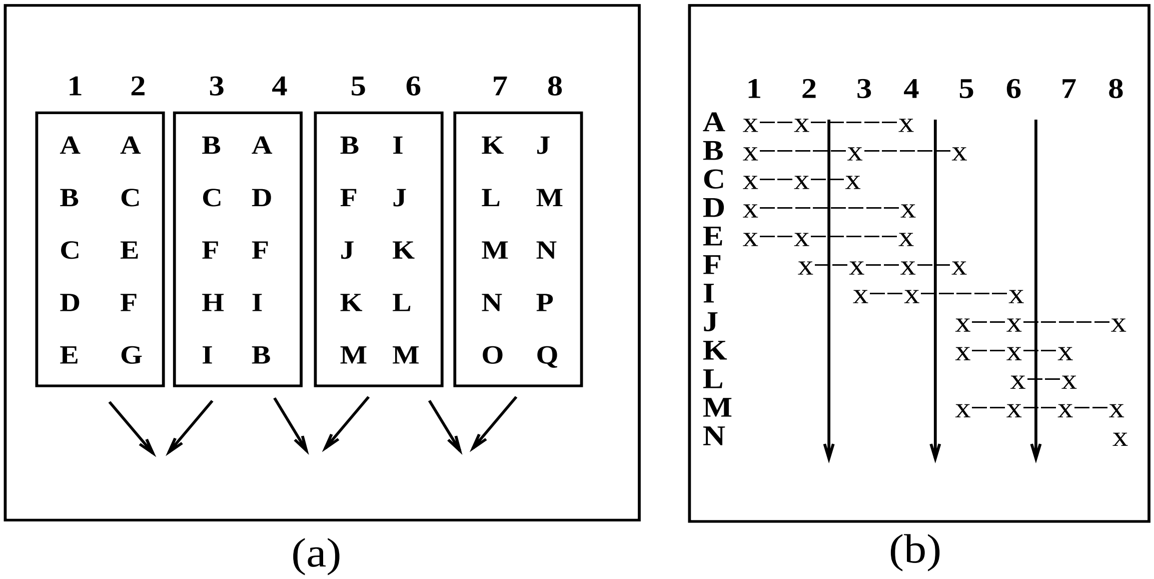


Figure 3: Illustration of the two proposed lexical cohesion comparison algorithms. Letters signify lexical items, numbers signify sentence numbers. (a) Similarity comparison of adjacent blocks with a window of size 2. Arrows indicate which blocks are compared to yield scores for sentence gaps 2, 4, and 6. Blocks would be shifted by one sentence for similarity measurements for gaps 3, 5, and 7. (b) Accumulating counts of chains of terms: 'x' indicates that the term occurs in the sentence, '-' indicates continuation of a chain, and arrows cut through the active chains that contribute to the cumulative count for sentence gaps 2, 4, and 6.

4.1 Reader Judgments

We obtained judgements from seven readers for each of thirteen magazine articles which satisfied the length criteria (between 1800 and 2500 words)[4]) and which contained little structural demarkation. The judges were asked simply to mark the paragraph boundaries at which the topic changed; they were not given more explicit instructions about the granularity of the segmentation.

Figure 4 shows the boundaries marked by seven judges on the *Stargazers* text. This format helps illuminate the general trends made by the judges and also helps show where and how often they disagree. For instance, all but one judge marked a boundary between paragraphs 2 and 3. The dissenting judge did mark a boundary after 3, as did two of the concurring judges. The next three major boundaries occur after paragraphs 5, 9, 12, and 13. There is some contention in the later paragraphs; three readers marked both 16 and 18, two marked 18 alone, and two marked 17 alone. The outline in Section 1 gives an idea of what each segment is about.

(Passonneau & Litman 1993) discuss at length the considerations that must go into the evaluation of segmentation algorithms according to reader judgement information. As Figure 4 shows, agreement among judges is not perfect, but trends can be discerned. In our evaluation we follow the suggestions of (Passonneau & Litman 1993); we determine which

[4] One longer text of 2932 words was used since reader judgements had been obtained for it from an earlier experiment. Note that this represents an amount of test data on the order of that used in the experiments of (Passonneau & Litman 1993). Judges were technical researchers. Two texts had three or four short headers which we removed.

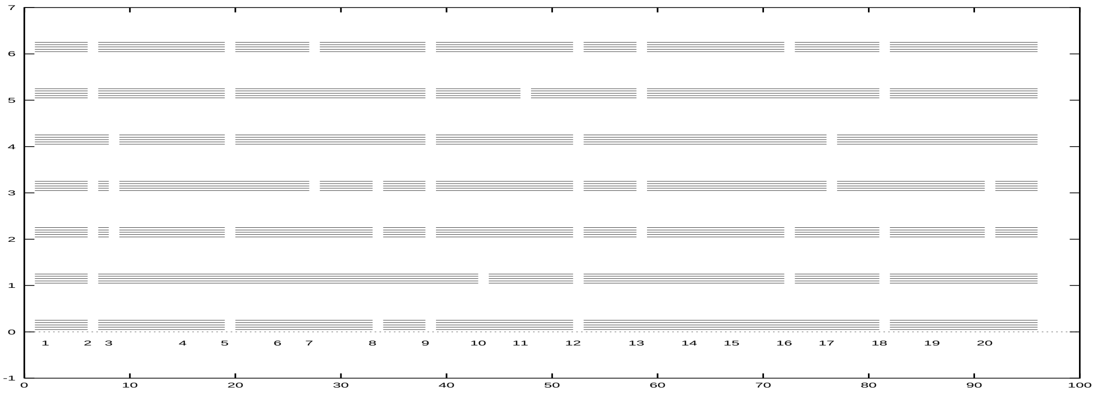


Figure 4: Judgements of seven readers on the *Stargazer* text. Internal numbers indicate location of gaps between paragraphs; x-axis indicates sentence gap number, y-axis indicates judge number, a break in a horizontal line indicates a judge-specified segment break.

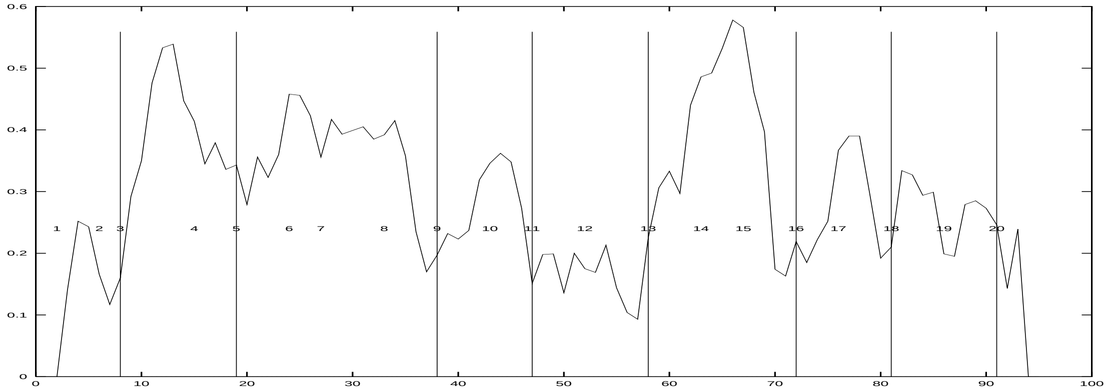


Figure 5: Results of the block similarity algorithm on the *Stargazer* text. Internal numbers indicate paragraph numbers, x-axis indicates sentence gap number, y-axis indicates similarity between blocks centered at the corresponding sentence gap. Vertical lines indicate boundaries chosen by the algorithm; for example, the leftmost vertical line represents a boundary after paragraph 3. Note how these align with the boundary gaps of Figure 4 above.

boundaries have been marked by a significant [5] proportion of the judges and label those paragraph gaps as "true" segment boundaries. The remaining gaps are considered non-boundaries. The algorithms are evaluated according to how many true boundaries they select out of the total selected (precision) and how many true boundaries are found out of the total possible (recall) (Salton 1988). The recall measure implicitly signals the number of missed boundaries (false negatives, or deletion errors); we also indicate the number of false positives, or insertion errors, explicitly.

[5] In Passonneau and Litman's data, if 4 or more out of 7 judges mark a boundary, the segmentation is found to be significant using a variation of the Q-test (Cochran 1950). We found similar results but included boundaries marked by only three judges, which in general better suited our task. Furthermore, for purposes of evaluation, paragraphs of three or fewer sentences were combined with their neighbor if that neighbor was deemed to be followed by a "true" boundary, as in paragraphs 2 and 3 of the *Stargazers* text.

	Precision		Recall	
	avg	sd	avg	sd
Baseline 33%	.44	.08	.37	.04
Baseline 41%	.43	.08	.42	.03
Chains	.64	.17	.58	.17
Blocks	.66	.18	.61	.13
Judges	.81	.06	.71	.06

Table 1: Precision and Recall values for 13 test texts.

4.2 Results

Figure 5 shows a plot of the results of applying the block comparison algorithm to the *Stargazer* text. When the lowermost portion of a valley is not located at a paragraph gap, the judgement is moved to the nearest paragraph gap.[6] For the most part, the regions of strong similarity correspond to the regions of strong agreement among the readers. (These results were fifth highest out of our 13 test texts.) Note however, that the similarity information around paragraph 12 is weak. This paragraph acts as a summary paragraph, summarizing the contents of the previous three and revisiting much of the terminology that occurred in them all in one location (in the spirit of a (Grosz & Sidner 1986) "pop" operation). Thus it displays low similarity both to itself and to its neighbors. This is an example of a breakdown caused by the assumption about the linear sequence of the subtopic discussions. It is possible that an additional pass through the text could be used to find structure of this kind.

The final paragraph is a summary of the entire text; the algorithm recognizes the change in terminology from the preceding paragraphs and marks a boundary; only two of the readers chose to differentiate the summary; for this reason the algorithm is judged to have made an error even though this sectioning decision is reasonable. This illustrates the inherent fallibility of testing against reader judgements, although in part this is because the judges were given loose constraints.

Following the advice of (Gale *et al.* 1992a), we compare our algorithm against both upper and lower bounds. The upper bound in this case is the reader judgement data. The lower bound is a baseline algorithm that is a simple, reasonable approach to the problem that can be automated. We determined from our test data that boundaries are placed in about 41% of the paragraph gaps, and wrote a program that places a boundary at each potential gap 41% of the time. We ran the program 10,000 times for each test text and computed precision and recall for the average of the results; these scores appear in Table 1 (results at 33% are also shown for comparison purposes).

Table 1 shows that the blocking algorithm is sandwiched between the upper and lower bounds. The block similarity algorithm seems to work slightly better than the chaining algorithm, although the difference may not prove significant over the long run. Table 2 shows some of these results in more detail.

In many cases the algorithms are almost correct but off by one paragraph, especially in the texts that the algorithm performs poorly on. When we allow the block similarity algorithm to be off by one paragraph, there is dramatic improvement in the scores for the texts that lower part of Table 2, yielding an overall precision of 83% and recall of 78%. As in Figure 5, we often see that where the algorithm is incorrect, e.g., paragraph gap 11, the overall blocking is very close to what the judges intended.

[6] This might be explained in part by (Stark 1988) who shows that readers disagree measurably about where to place paragraph boundaries when presented with texts with those boundaries removed.

	Total	Baseline 41% (avg)				Blocks				Chains				Judges (avg)			
Text	Possible	Prec	Rec	C	I	Prec	Rec	C	I	Prec	Rec	C	I	Prec	Rec	C	I
1	9	.45	.45	4	5	1.0	.78	7	0	1.0	.78	7	0	.78	.78	7	2
2	9	.50	.45	4	4	.88	.78	7	1	.75	.33	3	1	.88	.78	7	1
3	9	.40	.45	4	6	.78	.78	7	2	.56	.56	5	4	.75	.67	6	2
4	12	.63	.42	5	3	.86	.50	6	1	.56	.33	5	4	.90	.58	10	1
5	8	.43	.38	3	4	.70	.75	6	2	.86	.75	6	1	.86	.75	6	1
6	8	.40	.38	3	9	.60	.75	6	3	.42	.63	5	8	.75	.75	6	2
7	9	.36	.45	4	7	.60	.55	5	3	.40	.44	4	6	.75	.67	6	2
8	8	.43	.38	3	4	.50	.62	5	4	.67	.75	6	3	.86	.75	6	1
9	9	.36	.45	4	7	.50	.44	4	3	.60	.33	3	2	.75	.67	6	2
10	8	.50	.38	3	3	.50	.50	4	3	.63	.63	5	3	.86	.75	6	1
11	9	.36	.45	4	7	.50	.44	4	4	.71	.71	5	2	.75	.67	6	2
12	9	.45	.45	4	5	.50	.55	5	5	.54	.54	7	6	.86	.67	6	1
13	10	.36	.40	4	7	.30	.50	5	9	.60	.60	6	4	.78	.70	7	2

Table 2: Scores by text, showing precision and recall. (C) indicates the number of correctly placed boundaries, (I) indicates the number of inserted boundaries. The number of deleted boundaries can be determined by subtracting (C) from Total Possible.

5 Summary and Future Work

This paper has described algorithms for the segmentation of expository texts into discourse units that reflect the subtopic structure of expository text. We have introduced the notion of the recognition of multiple simultaneous themes as an approximation to Chafe's Flow Model of discourse and Skorodch'ko's text structure types. The algorithms are fully implemented: term repetition alone, without use of thesaural relations, knowledge bases, or inference mechanisms, works well for many of the texts we've experimented with. The structure it obtains is coarse-grained but generally reflects human judgement data.

We are experimenting with techniques to improve information retrieval from lengthy texts making use of multi-paragraph segmentation and have some preliminary positive results. We have implemented a prototype information retrieval interface that classifies the main topics of texts and will soon incorporate the subtopic structure information. However, to be more useful the segments should also be labeled according to what subtopic discussions they contain.

In earlier work (Hearst 1993) we incorporated thesaural information into our algorithms; surprisingly we've found in our latest experiments that this information degrades the performance. This could very well be due to inferior term categorizaton algorithms or categories that are too large; therefore we do not feel the issue is closed, and instead consider successful grouping of related words as future work. (Kozima 1993) has suggested using a (computationally expensive) semantic similarity metric to find similarity among terms within a small window of text (5 to 7 words). This work does not incorporate the notion of multiple simultaneous themes but instead just tries to find breaks in semantic similarity among a small number of terms. A good strategy may be to substitute this kind of similarity information for term repetition in algorithms like those described here.

The use of discourse cues for detection of segment boundaries and other discourse purposes has been extensively researched, although predominantly on spoken text (see (Hirschberg & Litman 1993) for a summary of six research groups' treatments of 64 cue words). It is possible that incorporation of such information may help improve the cases where the algorithm is off by one paragraph, as might reference resolution and an account of tense and aspect.

Acknowledgments

The author would like to thank Michael Schiff and Narciso Jaramillo for their helpful comments on this paper, and Robert Wilensky for his support of this line of research.

References

Brown, Gillian, & George Yule. 1983. *Discourse Analysis*. Cambridge Textbooks in Linguistics Series. Cambridge University Press.

Chafe, Wallace L. 1979. The flow of thought and the flow of language. In *Syntax and Semantics: Discourse and Syntax*, ed. by Talmy Givón, volume 12, 159–182. Academic Press.

Cochran, W. G. 1950. The comparison of percentages in matched samples. *Biometrika* 37.256–266.

Gale, William, Kenneth W. Church, & David Yarowsky. 1992a. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Meeting of the Association for Computational Linguistics*, 249–256.

Gale, William A., Kenneth W. Church, & David Yarowsky. 1992b. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 5-6.415–439.

Grosz, Barbara J., & Candace L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics* 12.172–204.

Halliday, M. A. K., & R. Hasan. 1976. *Cohesion in English*. London: Longman.

Hearst, Marti A. 1993. TextTiling: A quantitative approach to discourse segmentation. Technical Report 93/24, Sequoia 2000, University of California, Berkeley.

——, 1994. *Subtopic Structuring of Full-Length Documents*. University of California, Berkeley dissertation. In preparation.

——, & Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 59–68, Pittsburgh, PA.

Hirschberg, Julia, & Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19.501–530.

Kozima, Hideki. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, 286–288, Columbus, OH.

Longacre, R. E. 1979. The paragraph as a grammatical unit. In *Syntax and Semantics: Discourse and Syntax*, ed. by Talmy Givón, volume 12, 115–134. Academic Press.

Mann, William C., & Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS 87-190, ISI.

Mooney, David J., M. Sandra Carberry, & Kathleen F. McCoy. 1990. The generation of high-level structure for extended explanations. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, volume 2, 276–281, Helsinki.

MORRIS, JANE. 1988. Lexical cohesion, the thesaurus, and the structure of text. Technical Report CSRI-219, Computer Systems Research Institute, University of Toronto.

——, & GRAEME HIRST. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17.21–48.

PASSONNEAU, REBECCA J., & DIANE J. LITMAN. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 148–155.

SALTON, GERARD. 1988. *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

——, J. ALLAN, & CHRIS BUCKLEY. 1993. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 49–58, Pittsburgh, PA.

SCHÜTZE, HINRICH. 1993. Word space. In *Advances in Neural Information Processing Systems 5*, ed. by Stephen J. Hanson, Jack D. Cowan, & C. Lee Giles. San Mateo CA: Morgan Kaufmann.

SIBUN, PENELOPE. 1992. Generating text without trees. *Computational Intelligence: Special Issue on Natural Language Generation* 8.102–122.

SKOROCHOD'KO, E.F. 1972. Adaptive method of automatic abstracting and indexing. In *Information Processing 71: Proceedings of the IFIP Congress 71*, ed. by C.V. Freiman, 1179–1182. North-Holland Publishing Company.

STARK, HEATHER. 1988. What do paragraph markers do? *Discourse Processes* 11.275–304.

TANNEN, DEBORAH. 1989. *Talking Voices: Repetition, dialogue, and imagery in conversational discourse*. Studies in Interactional Sociolinguistics 6. Cambridge University Press.

WALKER, MARILYN. 1991. Redundancy in collaborative dialogue. In *AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, ed. by Julia Hirschberg, Diane Litman, Kathy McCoy, & Candy Sidner, Pacific Grove, CA.

YAROWSKY, DAVID. 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 454–460, Nantes, France.